Потоковое чтение csv

```
[student559_10@bigdataanalytics-worker-3 ~]$ hdfs dfs -mkdir for_stream
[student559_10@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 2 items
drwxr-xr-x   - student559_10 student559_10          0 2022-01-20 12:47 .sparkStaging
drwxr-xr-x   - student559_10 student559_10          0 2022-01-28 12:51 for_stream
[student559_10@bigdataanalytics-worker-3 ~]$
```

```
[student559_10@bigdataanalytics-worker-3 ~]$ cat for_stream/white_house_2017_salaries.csv
NAME,STATUS,SALARY,PAY BASIS,POSITION TITLE
"Alexander, Monica K.",Employee,"$56,000.00 ",Per Annum,EXECUTIVE ASSISTANT
"Ambrosini, Michael J.",Employee,"$95,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND DIRECTOR OF THE OFFICE OF THE CHIEF OF STAFF
"Amin, Stacy C.",Employee,"$140,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT
"Andersen, Whitney N.",Employee,"$94,000.00 ",Per Annum,DEPUTY DIRECTOR OF OPERATIONS FOR THE WHITE HOUSE MANAGEMENT OFFICE
"Anderson, Alexander J.",Employee,"$77,000.00 ",Per Annum,DIRECTOR OF DIGITAL ENGAGEMENT
"Angelson, Alexander J.",Employee,"$95,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT FOR LEGISLATIVE AFFAIRS
"Augustine, Rene I.",Employee,"$140,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND SENIOR ASSOCIATE COUNSEL TO THE PRESIDENT
"Baitel, Rachael",Employee,"$62,000.00 ",Per Annum,EXECUTIVE ASSISTANT
"Baldwin, Brittany L.",Employee,"$70,000.00 ",Per Annum,SPEECHWRITER
"Banks, George D.",Employee,"$140,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY
"Bannon, Stephen K.",Employee,"$179,700.00 ",Per Annum,ASSISTANT TO THE PRESIDENT AND CHIEF STRATEGIST AND SENIOR COUNSELOR
"Barger, Lara R.",Employee,"$83,000.00 ",Per Annum,SENIOR DIGITAL STRATEGIST
"Bash, John F.",Employee,"$130,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT
"Bash, Zina G.",Employee,"$140,000.00 ",Per Annum,"SPECIAL ASSISTANT TO THE PRESIDENT FOR REGULATORY REFORM, LEGAL AND IMMIGRATION POLICY"
"Beley, James P.",Employee,"$47,000.00 ",Per Annum,WRITER FOR CORRESPONDENCE
"Berkowitz, Avrahm J.",Employee,"$115,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND ASSISTANT TO THE SENIOR ADVISOR
"Biddle, Emily K.",Employee,"$70,000.00 ",Per Annum,DEPUTY SOCIAL SECRETARY
"Bis, Justin B.",Employee,"$62,000.00 ",Per Annum,DEPUTY ASSOCIATE DIRECTOR
"Blair, Patricia A.",Employee,"$102,212.00 ",Per Annum,CHIEF CALLIGRAPHER
"Blase, Brian C.",Employee,"$115,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY
"Block, Monica J.",Employee,"$140,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND DEPUTY DIRECTOR OF WHITE HOUSE MANAGEMENT AND ADMINISTRATION
"Blount, Mallory N.",Employee,"$40,000.00 ",Per Annum,ASSOCIATE DIRECTOR
"Blount, Patricia H.",Employee,"$61,829.00 ",Per Annum,RECORDS MANAGEMENT ANALYST
"Bock, Caroline E.",Employee,"$115,000.00 ",Per Annum,SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE DIRECTOR OF PRESIDENTIAL PERSONNEL
```

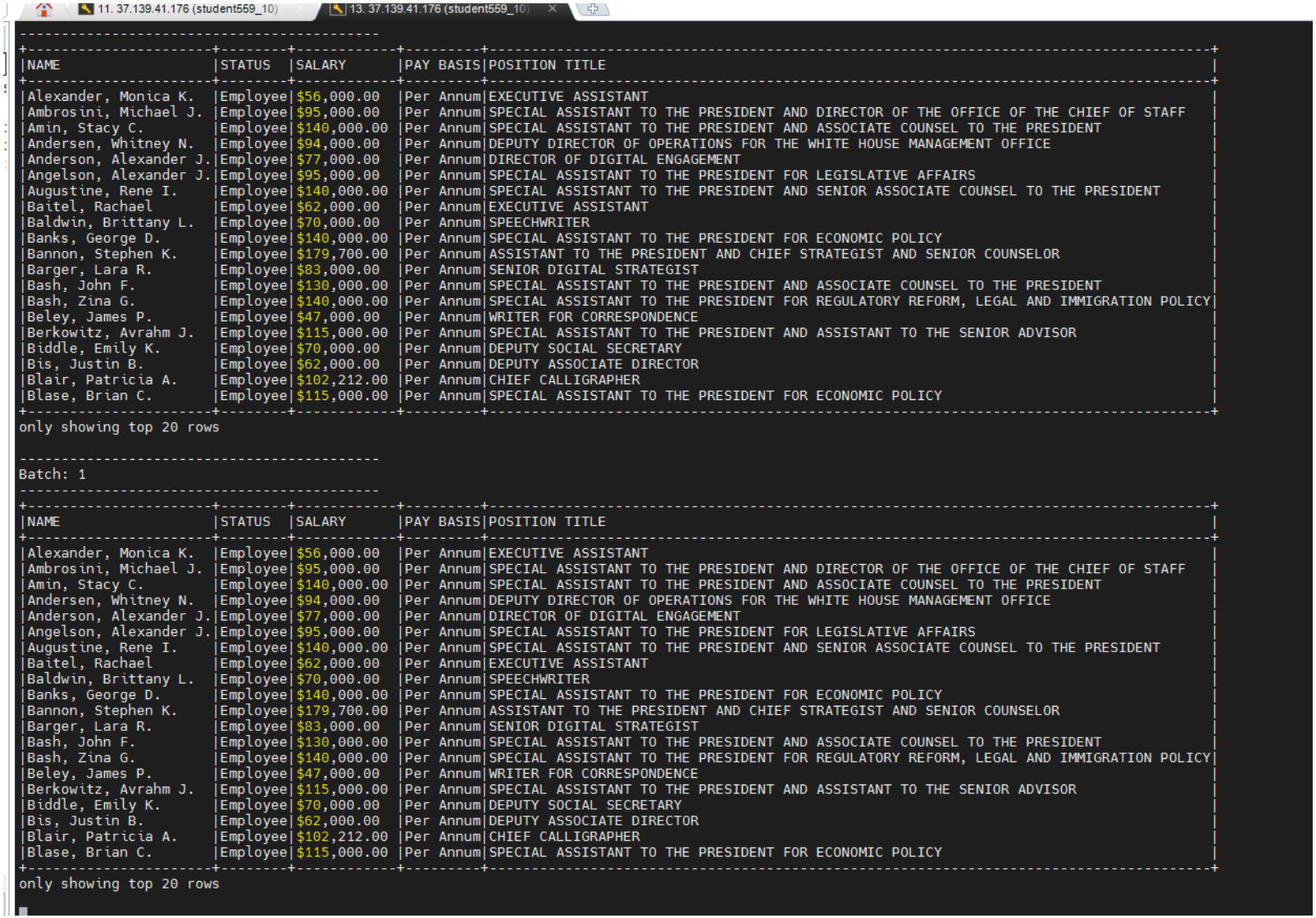
```
      /__/__  ___ _____/ /__
     _\ \/ _ \/ _ `/ __/  '_/
    /__ / .__/\_,_/_/ /_/\_\   version 2.3.2.3.1.4.0-315
       /_/

Using Python version 2.7.5 (default, Nov 16 2020 22:23:17)
SparkSession available as 'spark'.
>>> from pyspark.sql import functions as F
>>> from pyspark.sql.types import *
>>> def console_output(df, freq):
...     return df.writeStream \
...         .format("console") \
...         .trigger(processingTime='%s seconds' % freq ) \
...         .options(truncate=False) \
...         .start()
...
>>> schema = StructType() \
...     .add("NAME", StringType()) \
...     .add("STATUS", StringType()) \
...     .add("SALARY", FloatType()) \
...     .add("PAY BASIS", StringType()) \
...     .add("POSITION TITLE", StringType())
>>>
```

```
>>> raw_files = spark \
...     .readStream \
...     .format("csv") \
...     .schema(schema) \
...     .options(path="for_stream", header=True, sep = ",") \
...     .load()
```

```
[student559_10@bigdataanalytics-worker-3 ~]$ cd for_stream
[student559_10@bigdataanalytics-worker-3 for_stream]$ hdfs dfs -put white_house_2017_salaries.csv for_stream
[student559_10@bigdataanalytics-worker-3 for_stream]$
```

```
>>> out = console_output(raw_files, 10)
-------------------------------------------
Batch: 0
-------------------------------------------
+----------------------+--------+------------+---------+-----------------------------------------------------------------------------------------------+
|NAME                  |STATUS  |SALARY      |PAY BASIS|POSITION TITLE                                                                                 |
+----------------------+--------+------------+---------+-----------------------------------------------------------------------------------------------+
|Alexander, Monica K.  |Employee|$56,000.00  |Per Annum|EXECUTIVE ASSISTANT                                                                            |
|Ambrosini, Michael J. |Employee|$95,000.00  |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND DIRECTOR OF THE OFFICE OF THE CHIEF OF STAFF            |
|Amin, Stacy C.        |Employee|$140,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT                      |
|Andersen, Whitney N.  |Employee|$94,000.00  |Per Annum|DEPUTY DIRECTOR OF OPERATIONS FOR THE WHITE HOUSE MANAGEMENT OFFICE                            |
|Anderson, Alexander J.|Employee|$77,000.00  |Per Annum|DIRECTOR OF DIGITAL ENGAGEMENT                                                                 |
|Angelson, Alexander J.|Employee|$95,000.00  |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR LEGISLATIVE AFFAIRS                                     |
|Augustine, Rene I.    |Employee|$140,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND SENIOR ASSOCIATE COUNSEL TO THE PRESIDENT               |
|Baitel, Rachael       |Employee|$62,000.00  |Per Annum|EXECUTIVE ASSISTANT                                                                            |
|Baldwin, Brittany L.  |Employee|$70,000.00  |Per Annum|SPEECHWRITER                                                                                   |
|Banks, George D.      |Employee|$140,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY                                         |
|Bannon, Stephen K.    |Employee|$179,700.00 |Per Annum|ASSISTANT TO THE PRESIDENT AND CHIEF STRATEGIST AND SENIOR COUNSELOR                           |
|Barger, Lara R.       |Employee|$83,000.00  |Per Annum|SENIOR DIGITAL STRATEGIST                                                                      |
|Bash, John F.         |Employee|$130,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT                      |
|Bash, Zina G.         |Employee|$140,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR REGULATORY REFORM, LEGAL AND IMMIGRATION POLICY         |
|Beley, James P.       |Employee|$47,000.00  |Per Annum|WRITER FOR CORRESPONDENCE                                                                      |
|Berkowitz, Avrahm J.  |Employee|$115,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSISTANT TO THE SENIOR ADVISOR                         |
|Biddle, Emily K.      |Employee|$70,000.00  |Per Annum|DEPUTY SOCIAL SECRETARY                                                                        |
|Bis, Justin B.        |Employee|$62,000.00  |Per Annum|DEPUTY ASSOCIATE DIRECTOR                                                                      |
|Blair, Patricia A.    |Employee|$102,212.00 |Per Annum|CHIEF CALLIGRAPHER                                                                             |
|Blase, Brian C.       |Employee|$115,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY                                         |
+----------------------+--------+------------+---------+-----------------------------------------------------------------------------------------------+
only showing top 20 rows
```

```
[student559_10@bigdataanalytics-worker-3 for_stream]$ hdfs dfs -put white_house_2017_salaries.csv for_stream
[student559_10@bigdataanalytics-worker-3 for_stream]$ hdfs dfs -put white_house_2017_salaries_2.csv for_stream
[student559_10@bigdataanalytics-worker-3 for_stream]$
```



```
11. 37.139.41.176 (student559_10)   13. 37.139.41.176 (student559_10)
-------------------------------------------
+----------------------+--------+-----------+---------+-----------------------------------------------------------------------------------------------------+
|NAME                  |STATUS  |SALARY     |PAY BASIS|POSITION TITLE                                                                                       |
+----------------------+--------+-----------+---------+-----------------------------------------------------------------------------------------------------+
|Alexander, Monica K.  |Employee|$56,000.00 |Per Annum|EXECUTIVE ASSISTANT                                                                                  |
|Ambrosini, Michael J. |Employee|$95,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND DIRECTOR OF THE OFFICE OF THE CHIEF OF STAFF                  |
|Amin, Stacy C.        |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT                            |
|Andersen, Whitney N.  |Employee|$94,000.00 |Per Annum|DEPUTY DIRECTOR OF OPERATIONS FOR THE WHITE HOUSE MANAGEMENT OFFICE                                  |
|Anderson, Alexander J.|Employee|$77,000.00 |Per Annum|DIRECTOR OF DIGITAL ENGAGEMENT                                                                       |
|Angelson, Alexander J.|Employee|$95,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR LEGISLATIVE AFFAIRS                                           |
|Augustine, Rene I.    |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND SENIOR ASSOCIATE COUNSEL TO THE PRESIDENT                     |
|Baitel, Rachael       |Employee|$62,000.00 |Per Annum|EXECUTIVE ASSISTANT                                                                                  |
|Baldwin, Brittany L.  |Employee|$70,000.00 |Per Annum|SPEECHWRITER                                                                                         |
|Banks, George D.      |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY                                               |
|Bannon, Stephen K.    |Employee|$179,700.00|Per Annum|ASSISTANT TO THE PRESIDENT AND CHIEF STRATEGIST AND SENIOR COUNSELOR                                 |
|Barger, Lara R.       |Employee|$83,000.00 |Per Annum|SENIOR DIGITAL STRATEGIST                                                                            |
|Bash, John F.         |Employee|$130,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT                            |
|Bash, Zina G.         |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR REGULATORY REFORM, LEGAL AND IMMIGRATION POLICY               |
|Beley, James P.       |Employee|$47,000.00 |Per Annum|WRITER FOR CORRESPONDENCE                                                                            |
|Berkowitz, Avrahm J.  |Employee|$115,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSISTANT TO THE SENIOR ADVISOR                               |
|Biddle, Emily K.      |Employee|$70,000.00 |Per Annum|DEPUTY SOCIAL SECRETARY                                                                              |
|Bis, Justin B.        |Employee|$62,000.00 |Per Annum|DEPUTY ASSOCIATE DIRECTOR                                                                            |
|Blair, Patricia A.    |Employee|$102,212.00|Per Annum|CHIEF CALLIGRAPHER                                                                                   |
|Blase, Brian C.       |Employee|$115,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY                                               |
+----------------------+--------+-----------+---------+-----------------------------------------------------------------------------------------------------+
only showing top 20 rows

-------------------------------------------
Batch: 1
-------------------------------------------
+----------------------+--------+-----------+---------+-----------------------------------------------------------------------------------------------------+
|NAME                  |STATUS  |SALARY     |PAY BASIS|POSITION TITLE                                                                                       |
+----------------------+--------+-----------+---------+-----------------------------------------------------------------------------------------------------+
|Alexander, Monica K.  |Employee|$56,000.00 |Per Annum|EXECUTIVE ASSISTANT                                                                                  |
|Ambrosini, Michael J. |Employee|$95,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND DIRECTOR OF THE OFFICE OF THE CHIEF OF STAFF                  |
|Amin, Stacy C.        |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT                            |
|Andersen, Whitney N.  |Employee|$94,000.00 |Per Annum|DEPUTY DIRECTOR OF OPERATIONS FOR THE WHITE HOUSE MANAGEMENT OFFICE                                  |
|Anderson, Alexander J.|Employee|$77,000.00 |Per Annum|DIRECTOR OF DIGITAL ENGAGEMENT                                                                       |
|Angelson, Alexander J.|Employee|$95,000.00 |Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR LEGISLATIVE AFFAIRS                                           |
|Augustine, Rene I.    |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND SENIOR ASSOCIATE COUNSEL TO THE PRESIDENT                     |
|Baitel, Rachael       |Employee|$62,000.00 |Per Annum|EXECUTIVE ASSISTANT                                                                                  |
|Baldwin, Brittany L.  |Employee|$70,000.00 |Per Annum|SPEECHWRITER                                                                                         |
|Banks, George D.      |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY                                               |
|Bannon, Stephen K.    |Employee|$179,700.00|Per Annum|ASSISTANT TO THE PRESIDENT AND CHIEF STRATEGIST AND SENIOR COUNSELOR                                 |
|Barger, Lara R.       |Employee|$83,000.00 |Per Annum|SENIOR DIGITAL STRATEGIST                                                                            |
|Bash, John F.         |Employee|$130,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSOCIATE COUNSEL TO THE PRESIDENT                            |
|Bash, Zina G.         |Employee|$140,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR REGULATORY REFORM, LEGAL AND IMMIGRATION POLICY               |
|Beley, James P.       |Employee|$47,000.00 |Per Annum|WRITER FOR CORRESPONDENCE                                                                            |
|Berkowitz, Avrahm J.  |Employee|$115,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT AND ASSISTANT TO THE SENIOR ADVISOR                               |
|Biddle, Emily K.      |Employee|$70,000.00 |Per Annum|DEPUTY SOCIAL SECRETARY                                                                              |
|Bis, Justin B.        |Employee|$62,000.00 |Per Annum|DEPUTY ASSOCIATE DIRECTOR                                                                            |
|Blair, Patricia A.    |Employee|$102,212.00|Per Annum|CHIEF CALLIGRAPHER                                                                                   |
|Blase, Brian C.       |Employee|$115,000.00|Per Annum|SPECIAL ASSISTANT TO THE PRESIDENT FOR ECONOMIC POLICY                                               |
+----------------------+--------+-----------+---------+-----------------------------------------------------------------------------------------------------+
only showing top 20 rows
```

```
>>> extra_files = raw_files[raw_files["POSITION TITLE"].isin("STAFF ASSISTANT")]
```

```
out = console_output(extra_files, 10)
-------------------------------------------
Batch: 0
-------------------------------------------
+----------------------+--------+-----------+---------+---------------+
|NAME                  |STATUS  |SALARY     |PAY BASIS|POSITION TITLE |
+----------------------+--------+-----------+---------+---------------+
|Cabaniss, Anna K.     |Employee|$42,000.00 |Per Annum|STAFF ASSISTANT|
|Carroll, III, James W.|Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Cheap, Casey C.       |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|David, Blandon J.     |Employee|$56,000.00 |Per Annum|STAFF ASSISTANT|
|Finzer, Mary C.       |Employee|$56,000.00 |Per Annum|STAFF ASSISTANT|
|Hennessey, Millicent S.|Employee|$40,000.00 |Per Annum|STAFF ASSISTANT|
|Mahfouz, Michael D.   |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|McAvoy, Ryan P.       |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Mitchelson, William J.|Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Pedersen, Brittany N. |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Redle, Alexander J.   |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Reese, Christopher M. |Employee|$62,000.00 |Per Annum|STAFF ASSISTANT|
|Sewell, John B.       |Employee|$42,000.00 |Per Annum|STAFF ASSISTANT|
|Szabo, Thomas G.      |Employee|$51,000.00 |Per Annum|STAFF ASSISTANT|
|Teresa, Tyler C.      |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Veletsis, Alexandra E.|Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Weber, Lauren F.      |Employee|$56,000.00 |Per Annum|STAFF ASSISTANT|
|Zager, Samantha L.    |Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
|Cabaniss, Anna K.     |Employee|$42,000.00 |Per Annum|STAFF ASSISTANT|
|Carroll, III, James W.|Employee|$47,000.00 |Per Annum|STAFF ASSISTANT|
+----------------------+--------+-----------+---------+---------------+
only showing top 20 rows
```

Потоковое чтение из Kafka

```
Last login: Fri Jan 28 14:01:45 2022 from broadband-46-242-8-209.ip.moscow.rt.ru
[student559_10@bigdataanalytics-worker-3 ~]$ cd for_stream
[student559_10@bigdataanalytics-worker-3 for_stream]$ cat iris.json
[
  {"sepalLength": 5.1, "sepalWidth": 3.5, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.9, "sepalWidth": 3.0, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.7, "sepalWidth": 3.2, "petalLength": 1.3, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.6, "sepalWidth": 3.1, "petalLength": 1.5, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.0, "sepalWidth": 3.6, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.4, "sepalWidth": 3.9, "petalLength": 1.7, "petalWidth": 0.4, "species": "setosa"},
  {"sepalLength": 4.6, "sepalWidth": 3.4, "petalLength": 1.4, "petalWidth": 0.3, "species": "setosa"},
  {"sepalLength": 5.0, "sepalWidth": 3.4, "petalLength": 1.5, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.4, "sepalWidth": 2.9, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.9, "sepalWidth": 3.1, "petalLength": 1.5, "petalWidth": 0.1, "species": "setosa"},
  {"sepalLength": 5.4, "sepalWidth": 3.7, "petalLength": 1.5, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.8, "sepalWidth": 3.4, "petalLength": 1.6, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.8, "sepalWidth": 3.0, "petalLength": 1.4, "petalWidth": 0.1, "species": "setosa"},
  {"sepalLength": 4.3, "sepalWidth": 3.0, "petalLength": 1.1, "petalWidth": 0.1, "species": "setosa"},
  {"sepalLength": 5.8, "sepalWidth": 4.0, "petalLength": 1.2, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.7, "sepalWidth": 4.4, "petalLength": 1.5, "petalWidth": 0.4, "species": "setosa"},
  {"sepalLength": 5.4, "sepalWidth": 3.9, "petalLength": 1.3, "petalWidth": 0.4, "species": "setosa"},
  {"sepalLength": 5.1, "sepalWidth": 3.5, "petalLength": 1.4, "petalWidth": 0.3, "species": "setosa"},
  {"sepalLength": 5.7, "sepalWidth": 3.8, "petalLength": 1.7, "petalWidth": 0.3, "species": "setosa"},
  {"sepalLength": 5.1, "sepalWidth": 3.8, "petalLength": 1.5, "petalWidth": 0.3, "species": "setosa"},
  {"sepalLength": 5.4, "sepalWidth": 3.4, "petalLength": 1.7, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.1, "sepalWidth": 3.7, "petalLength": 1.5, "petalWidth": 0.4, "species": "setosa"}
```

```
[student559_10@bigdataanalytics-worker-3 for_stream]$ /usr/hdp/3.1.4.0-315/kafka/bin/kafka-console-consumer.sh --topic test_lesson_3_sapr --bootstrap-server bigdataanalytics-worker-3:6667
        [
  {"sepalLength": 5.1, "sepalWidth": 3.5, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.9, "sepalWidth": 3.0, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.7, "sepalWidth": 3.2, "petalLength": 1.3, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.6, "sepalWidth": 3.1, "petalLength": 1.5, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.0, "sepalWidth": 3.6, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.4, "sepalWidth": 3.9, "petalLength": 1.7, "petalWidth": 0.4, "species": "setosa"},
  {"sepalLength": 4.6, "sepalWidth": 3.4, "petalLength": 1.4, "petalWidth": 0.3, "species": "setosa"},
  {"sepalLength": 5.0, "sepalWidth": 3.4, "petalLength": 1.5, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.4, "sepalWidth": 2.9, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.9, "sepalWidth": 3.1, "petalLength": 1.5, "petalWidth": 0.1, "species": "setosa"},
  {"sepalLength": 5.4, "sepalWidth": 3.7, "petalLength": 1.5, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.8, "sepalWidth": 3.4, "petalLength": 1.6, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 4.8, "sepalWidth": 3.0, "petalLength": 1.4, "petalWidth": 0.1, "species": "setosa"},
  {"sepalLength": 4.3, "sepalWidth": 3.0, "petalLength": 1.1, "petalWidth": 0.1, "species": "setosa"},
  {"sepalLength": 5.8, "sepalWidth": 4.0, "petalLength": 1.2, "petalWidth": 0.2, "species": "setosa"},
  {"sepalLength": 5.7, "sepalWidth": 4.4, "petalLength": 1.5, "petalWidth": 0.4, "species": "setosa"},
```

```
        at org.apache.spark.sql.Dataset$$anonfun$head$1.apply(Dataset.scala:2489)
        at org.apache.spark.sql.Dataset$$anonfun$head$1.apply(Dataset.scala:2489)
        at org.apache.spark.sql.Dataset$$anonfun$52.apply(Dataset.scala:3259)
        at org.apache.spark.sql.execution.SQLExecution$.withNewExecutionId(SQLExecution.scala:77)
        at org.apache.spark.sql.Dataset.withAction(Dataset.scala:3258)
        at org.apache.spark.sql.Dataset.head(Dataset.scala:2489)
        at org.apache.spark.sql.Dataset.take(Dataset.scala:2703)
        at org.apache.spark.sql.Dataset.showString(Dataset.scala:254)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
        at py4j.Gateway.invoke(Gateway.java:282)
        at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:132)
        at py4j.commands.CallCommand.execute(CallCommand.java:79)
        at py4j.GatewayConnection.run(GatewayConnection.java:238)
        at java.lang.Thread.run(Thread.java:748)

>>>
>>> /usr/hdp/current/kafka-broker/bin/kafka-console-producer.sh --topic test_lesson_3_sapr --broker-list bigdataanalytics-worker-3:6667 < iris.json
  File "<stdin>", line 1
    /usr/hdp/current/kafka-broker/bin/kafka-console-producer.sh --topic test_lesson_3_sapr --broker-list bigdataanalytics-worker-3:6667 < iris.json
    ^
SyntaxError: invalid syntax
>>> exit()
[student559_10@bigdataanalytics-worker-3 for_stream]$
[student559_10@bigdataanalytics-worker-3 for_stream]$ /usr/hdp/current/kafka-broker/bin/kafka-console-producer.sh --topic test_lesson_3_sapr --broker-list bigdataanalytics-worker-3:6667 < iris.json
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>[student559_10@bigdataanalytics-worker-3 for_stream]$
```

```
[student559_10@bigdataanalytics-worker-3 for_stream]$  export SPARK_KAFKA_VERSION=0.10
[student559_10@bigdataanalytics-worker-3 for_stream]$
[student559_10@bigdataanalytics-worker-3 for_stream]$ pyspark --master local[1] --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.2
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.7.5 (default, Nov 16 2020, 22:23:17)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Ivy Default Cache set to: /home/student559_10/.ivy2/cache
The jars for the packages stored in: /home/student559_10/.ivy2/jars
:: loading settings :: url = jar:file:/usr/hdp/3.1.4.0-315/spark2/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-1c9aea20-e8d0-4bc4-961f-1ee6b0ef00ec;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.11;2.3.2 in central
        found org.apache.kafka#kafka-clients;0.10.0.1 in central
        found net.jpountz.lz4#lz4;1.3.0 in central
        found org.xerial.snappy#snappy-java;1.1.2.6 in central
        found org.slf4j#slf4j-api;1.7.16 in central
        found org.spark-project.spark#unused;1.0.0 in central
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.11/2.3.2/spark-sql-kafka-0-10_2.11-2.3.2.jar ...
        [SUCCESSFUL ] org.apache.spark#spark-sql-kafka-0-10_2.11;2.3.2!spark-sql-kafka-0-10_2.11.jar (121ms)
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/0.10.0.1/kafka-clients-0.10.0.1.jar ...
        [SUCCESSFUL ] org.apache.kafka#kafka-clients;0.10.0.1!kafka-clients.jar (79ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.0.0.jar ...
        [SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (36ms)
downloading https://repo1.maven.org/maven2/net/jpountz/lz4/lz4/1.3.0/lz4-1.3.0.jar ...
        [SUCCESSFUL ] net.jpountz.lz4#lz4;1.3.0!lz4.jar (50ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.2.6/snappy-java-1.1.2.6.jar ...
        [SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.2.6!snappy-java.jar(bundle) (87ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.16/slf4j-api-1.7.16.jar ...
        [SUCCESSFUL ] org.slf4j#slf4j-api;1.7.16!slf4j-api.jar (44ms)
:: resolution report :: resolve 4084ms :: artifacts dl 423ms
        :: modules in use:
        net.jpountz.lz4#lz4;1.3.0 from central in [default]
        org.apache.kafka#kafka-clients;0.10.0.1 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.11;2.3.2 from central in [default]
        org.slf4j#slf4j-api;1.7.16 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.2.6 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   6   |   6   |   6   |   0   ||   6   |   6   |
        ---------------------------------------------------------------------
```

```
>>> from pyspark.sql import functions as F
>>> from pyspark.sql.types import *
>>> kafka_brokers = "bigdataanalytics-worker-3:6667"
>>> kafka_topic = "test_lesson_3_sapr"
>>> def console_output(df, freq):
...     return df.writeStream \
...         .format("console") \
...         .trigger(processingTime='%s seconds' % freq ) \
...         .options(truncate=False) \
...         .start()
...
>>> raw_data = spark.read. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "test_lesson_3_sapr"). \
...     option("startingOffsets", "earliest"). \
...     option("endingOffsets", """{"test_lesson_3_sapr":{"0":20}}"""). \
...     load()

>>>
```

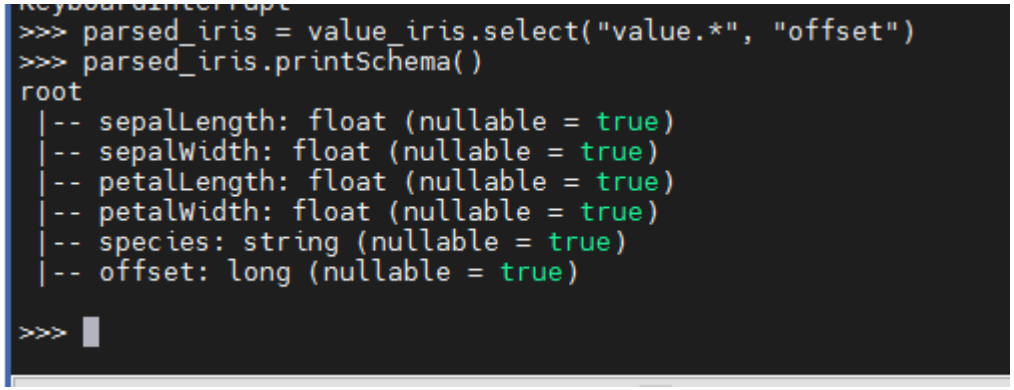
```
>>> raw_data.show(100)
+----+--------------------+----------------+---------+------+--------------------+-------------+
| key|               value|           topic|partition|offset|           timestamp|timestampType|
+----+--------------------+----------------+---------+------+--------------------+-------------+
|null|                [5B]|test_lesson_3_sapr|        0|     0|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     1|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     2|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     3|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     4|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     5|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     6|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     7|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     8|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|     9|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    10|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    11|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    12|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    13|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    14|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    15|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    16|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    17|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    18|2022-01-28 22:17:...|            0|
|null|[20 20 7B 22 73 6...|test_lesson_3_sapr|        0|    19|2022-01-28 22:17:...|            0|
+----+--------------------+----------------+---------+------+--------------------+-------------+

>>>
```

```
[student559_10@bigdataanalytics-worker-3 for_stream]$ /usr/hdp/3.1.4.0-315/kafka/bin/kafka-console-consumer.sh --topic test_lesson_3_sapr --bootstrap-server bigdataanalytics-worker-3:6667 --max-messages 2 --from-beginning
[
  {"sepalLength": 5.1, "sepalWidth": 3.5, "petalLength": 1.4, "petalWidth": 0.2, "species": "setosa"},
Processed a total of 2 messages
[student559_10@bigdataanalytics-worker-3 for_stream]$
```

```
>>> out = console_output(raw_data, 10)
-------------------------------------------
Batch: 0
-------------------------------------------
+----+------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------
----+------------+---------+------+-----------------------+-------------+
|key |value


    |topic       |partition|offset|timestamp              |timestampType|
+----+------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------
----+------------+---------+------+-----------------------+-------------+
|null|[5B]


    |shadrin_iris|0        |0     |2022-01-24 18:00:45.857|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 35 2E 31 2C 20 22 73 65 70 61 6C 57 69
74 68 22 3A 20 33 2E 35 2C 20 22 70 65 74 61 6C 4C 65 6E 67 74 68 22 3A 20 31 2E 34 2C 20 22 70 65 7
1 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73 61 22
 2C]|shadrin_iris|0        |1     |2022-01-24 18:00:45.863|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 34 2E 39 2C 20 22 73 65 70 61 6C 57 69
74 68 22 3A 20 33 2E 30 2C 20 22 70 65 74 61 6C 4C 65 6E 67 74 68 22 3A 20 31 2E 34 2C 20 22 70 65 7
1 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73 61 22
 2C]|shadrin_iris|0        |2     |2022-01-24 18:00:45.864|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 34 2E 37 2C 20 22 73 65 70 61 6C 57 69
74 68 22 3A 20 33 2E 32 2C 20 22 70 65 74 61 6C 4C 65 6E 67 74 68 22 3A 20 31 2E 33 2C 20 22 70 65 7
1 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73 61 22
 2C]|shadrin_iris|0        |3     |2022-01-24 18:00:45.864|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 34 2E 36 2C 20 22 73 65 70 61 6C 57 69
74 68 22 3A 20 33 2E 31 2C 20 22 70 65 74 61 6C 4C 65 6E 67 74 68 22 3A 20 31 2E 35 2C 20 22 70 65 7
1 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73 61 22
```

```
>>> raw_data.printSchema()
root
 |-- key: binary (nullable = true)
 |-- value: binary (nullable = true)
 |-- topic: string (nullable = true)
 |-- partition: integer (nullable = true)
 |-- offset: long (nullable = true)
 |-- timestamp: timestamp (nullable = true)
 |-- timestampType: integer (nullable = true)

>>>
```

```
>>> schema = StructType() \
...     .add("sepalLength", FloatType()) \
...     .add("sepalWidth", FloatType()) \
...     .add("petalLength", FloatType()) \
...     .add("petalWidth", FloatType()) \
...     .add("species", StringType())
>>> value_iris = raw_data \
...     .select(
...         F.from_json(F.col("value").cast("String"), schema).alias("value"),
...         "offset"
...     )
>>> value_iris.printSchema()
root
 |-- value: struct (nullable = true)
 |    |-- sepalLength: float (nullable = true)
 |    |-- sepalWidth: float (nullable = true)
 |    |-- petalLength: float (nullable = true)
 |    |-- petalWidth: float (nullable = true)
 |    |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
```


KeyboardInterrupt
>>> parsed_iris = value_iris.select("value.*", "offset")
>>> parsed_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
>>>

```
KeyboardInterrupt
>>> out = console_output(parsed_iris.withColumn(
...     'foo',
...     F.col("sepalLength") / F.col("petalLength")
... ), 30)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+-----------+----------+-----------+----------+-------+------+------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|foo               |
+-----------+----------+-----------+----------+-------+------+------------------+
|null       |null      |null       |null      |null   |0     |null              |
|5.1        |3.5       |1.4        |0.2       |setosa |1     |3.642857136775036 |
|4.9        |3.0       |1.4        |0.2       |setosa |2     |3.500000127724241 |
|4.7        |3.2       |1.3        |0.2       |setosa |3     |3.6153846012770066|
|4.6        |3.1       |1.5        |0.2       |setosa |4     |3.066666603088379 |
+-----------+----------+-----------+----------+-------+------+------------------+


-------------------------------------------
Batch: 1
-------------------------------------------
+-----------+----------+-----------+----------+-------+------+------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|foo               |
+-----------+----------+-----------+----------+-------+------+------------------+
|5.0        |3.6       |1.4        |0.2       |setosa |5     |3.5714286322496385|
|5.4        |3.9       |1.7        |0.4       |setosa |6     |3.176470555236184 |
|4.6        |3.4       |1.4        |0.3       |setosa |7     |3.2857142735500724|
|5.0        |3.4       |1.5        |0.2       |setosa |8     |3.3333333333333335|
|4.4        |2.9       |1.4        |0.2       |setosa |9     |3.142857264499277 |
+-----------+----------+-----------+----------+-------+------+------------------+

out.stop()
>>> out.stop()
```